

Applications Of The Divergence Theorem In Bayesian Inference And MaxEnt

Sergio Davis^{1,a)} and Gonzalo Gutiérrez^{2,b)}

¹*Comisión Chilena de Energía Nuclear, Casilla 188-D, Santiago, Chile.*

²*Grupo de Nanomateriales, Departamento de Física, Facultad de Ciencias, Universidad de Chile.*

^{a)}Corresponding author: sdavis@cchen.cl

^{b)}gonzalo@fisica.ciencias.uchile.cl

Abstract. Given a probability density $P(\mathbf{x}|\lambda)$, where \mathbf{x} represents continuous degrees of freedom and λ a set of parameters, it is possible to construct a general identity relating expectations of observable quantities, which is a generalization of the equipartition theorem in Thermodynamics.

In this work we explore some of the consequences of this relation, both in the context of sampling distributions and Bayesian posteriors, and how it can be used to extract some information without the need for explicit calculation of the partition function (or the Bayesian evidence, in the case of posterior expectations). Together with the general family of fluctuation theorems it constitutes a powerful tool for Bayesian/MaxEnt problems.

INTRODUCTION

In Statistical Thermodynamics, a system composed of N particles and connected to a thermal bath at temperature T is described by the *canonical ensemble*,

$$P(\mathbf{\Gamma}|\beta) = \frac{1}{Z(\beta)} \exp(-\beta\mathcal{H}(\mathbf{\Gamma})), \quad (1)$$

where $\beta = 1/k_B T$, the vector $\mathbf{\Gamma} = (x_1, x_2, \dots, x_{3N}, p_1, p_2, \dots, p_{3N})$ represents the $6N$ degrees of freedom of the system and \mathcal{H} is the Hamiltonian function, usually of the form

$$\mathcal{H}(\mathbf{\Gamma}) = \sum_{i=1}^N \frac{\mathbf{p}_i^2}{2m_i} + \Phi(x_1, \dots, x_{3N}). \quad (2)$$

Here β plays two roles: on the one hand it is the Lagrange multiplier which describes the probability distribution; on the other, it is associated with the average kinetic energy of the particles via the *equipartition theorem*,

$$\frac{3Nk_B T}{2} = \left\langle \sum_{i=1}^N \frac{\mathbf{p}_i^2}{2m_i} \right\rangle_\beta \quad (3)$$

where the average $\langle \cdot \rangle_\beta$ is taken using the canonical distribution in Eq. 1. This equipartition theorem, in a slightly more general form [1],

$$\beta \left\langle \Gamma_j \frac{\partial \mathcal{H}}{\partial \Gamma_i} \right\rangle_\beta = \delta_{ij}, \quad (4)$$

is obtained using the divergence (Gauss) theorem,

$$\int_{\Sigma} d\Sigma (u \mathbf{v} \cdot \hat{\mathbf{n}}) = \int_V dV (u \nabla \cdot \mathbf{v} + \mathbf{v} \cdot \nabla u), \quad (5)$$

where V is a volume in \mathbb{R}^{6N} representing a region of phase-space, Σ is a surface acting as a boundary of the volume V and \hat{n} is the unit vector normal to Σ .

Now, since the seminal paper by Jaynes [2] we know that Statistical Mechanics and the rule for constructing the canonical ensemble is just a particular application of the maximum entropy (MaxEnt) principle to physics; this suggests that some general form of the equipartition theorem holds for any continuous MaxEnt model fulfilling the conditions of the divergence theorem.

From now on we will work in general terms with a system described by N continuous degrees of freedom denoted collectively by $\mathbf{x} = (x_1, \dots, x_N)$. This system needs not be a “physical” system as in the case of the canonical ensemble before, but can be any set of N unknown quantities over which we do maximum entropy inference. We seek the optimal probability distribution consistent with m expectation constraints

$$\langle f_j(\mathbf{x}) \rangle = F_j, \quad (6)$$

when starting from the *prior*¹ distribution $P_0(\mathbf{x})$. The solution to this problem is the MaxEnt probability

$$P(\mathbf{x}|\lambda) = \frac{1}{Z(\lambda)} P_0(\mathbf{x}) \exp \left(- \sum_{j=1}^m \lambda_j f_j(\mathbf{x}) \right). \quad (7)$$

For this probability the authors have recently presented a generalization of the equipartition theorem [4],

$$\left\langle \frac{\partial}{\partial x_i} \omega(\mathbf{x}) \right\rangle_\lambda + \left\langle \omega(\mathbf{x}) \frac{\partial}{\partial x_i} \ln P_0(\mathbf{x}) \right\rangle_\lambda = \sum_{j=1}^m \lambda_j \left\langle \omega(\mathbf{x}) \frac{\partial}{\partial x_i} f_j(\mathbf{x}) \right\rangle_\lambda. \quad (8)$$

where $\omega(\mathbf{x})$ is an arbitrary, differentiable function of the coordinates \mathbf{x} .

This theorem relates the constraining functions f_j with their conjugate² Lagrange multipliers λ_j and was thus named the *conjugate variables theorem*, a neutral name which attempts to avoid any connection with the physical idea of equipartition (which might also be confused with the idea of “asymptotic equipartition” already present in Information Theory[5]).

In this work we start by providing a short proof of the CVT, which is more clear and general than the one in Ref. [4], and then review some applications to Statistical Mechanics, dynamical systems and Bayesian inference.

Variation inside the expectation: Conjugate variables theorem

Consider a model $P(\mathbf{x}|\lambda)$ where $\mathbf{x} = (x_1, \dots, x_N)$ represents a vector of random (or unknown) continuous variables which describe a system, and $\lambda = (\lambda_1, \dots, \lambda_m)$ denotes the vector of parameters of the model. We take every state \mathbf{x} as a point in \mathbb{R}^N .

We construct the expectation of a partial derivative $\partial\omega/\partial x_i$ of an arbitrary function $\omega(\mathbf{x}, \lambda)$ of the state and the parameters, and rewrite it using the divergence theorem (Equation 5),

$$\left\langle \frac{\partial \omega}{\partial x_i} \right\rangle_\lambda = \int d\mathbf{x} P(\mathbf{x}|\lambda) \frac{\partial}{\partial x_i} \omega(\mathbf{x}, \lambda) = \int d\Sigma (\hat{\mathbf{e}}_i \cdot \hat{\mathbf{n}}) P(\mathbf{x}|\lambda) \omega(\mathbf{x}, \lambda) - \int d\mathbf{x} \omega(\mathbf{x}, \lambda) \frac{\partial}{\partial x_i} P(\mathbf{x}|\lambda) \quad (9)$$

where we have chosen $u = P(\mathbf{x}|\lambda)$ and $\mathbf{v} = \hat{\mathbf{e}}_i \omega(\mathbf{x}, \lambda)$. As $P(\mathbf{x}|\lambda)$ must be properly normalized, it must vanish at $x_i = \pm\infty$ so the surface term also vanishes. Then, the second term can also be written as an expectation, and we finally obtain the identity

$$\left\langle \frac{\partial \omega}{\partial x_i} \right\rangle_\lambda = - \left\langle \omega \frac{\partial}{\partial x_i} \ln P(\mathbf{x}|\lambda) \right\rangle_\lambda, \quad (10)$$

valid for an arbitrary, differentiable function ω . This is the most general form of the CVT for any continuous model $P(\mathbf{x}|\lambda)$. The particular case in Eq. 8 is obtained by replacing the MaxEnt probability distribution given by Eq. 7.

¹ Jaynes, in the original formulation of MaxEnt, takes P_0 to be the invariant measure of the state space. However, recently [3] it has been shown that this is too restrictive: MaxEnt can be formulated as a method of updating an arbitrary *a priori* distribution under constraints.

² Conjugate in the sense of Thermodynamics.

A direct consequence of this identity is readily seen by setting $\omega = 1$, it shows that the condition for the probability maximum is also valid in expectation over all states, that is,

$$\left\langle \frac{\partial}{\partial x_i} \ln P(\mathbf{x}|\lambda) \right\rangle_\lambda = 0. \quad (11)$$

Variation outside the expectation: Fluctuation theorems

Under the same conditions of the CVT, let us now take the derivative of the expectation $\langle \omega(\mathbf{x}, \lambda) \rangle_\lambda$ with respect to one of the parameters, λ_j ,

$$\frac{\partial}{\partial \lambda_j} \langle \omega \rangle_\lambda = \frac{\partial}{\partial \lambda_j} \left(\int d\mathbf{x} \omega(\mathbf{x}, \lambda) P(\mathbf{x}|\lambda) \right) = \int d\mathbf{x} \left(P(\mathbf{x}|\lambda) \frac{\partial \omega}{\partial \lambda_j} + \omega \frac{\partial}{\partial \lambda_j} P(\mathbf{x}|\lambda) \right). \quad (12)$$

Again, rewriting the second term as an expectation, we have

$$\frac{\partial}{\partial \lambda_j} \langle \omega \rangle_\lambda = \left\langle \frac{\partial \omega}{\partial \lambda_j} \right\rangle_\lambda + \left\langle \omega \frac{\partial}{\partial \lambda_j} \ln P(\mathbf{x}|\lambda) \right\rangle_\lambda, \quad (13)$$

which is similar to the CVT in that we have a “free” function $\omega(x, \lambda)$. This is the most general version of a family of “fluctuation theorems”. The particular case where $P(\mathbf{x}|\lambda)$ is a Maximum Entropy model was already given by Jaynes in his book[6],

$$\frac{\partial}{\partial \lambda_j} \langle \omega \rangle_\lambda = - \left\langle \delta \omega \delta f_j \right\rangle_\lambda. \quad (14)$$

Here the right-hand side is in fact a covariance or correlation between fluctuations. In particular, in the canonical ensemble the following relation connecting the energy fluctuations with the heat capacity [7],

$$\left\langle (\mathcal{H} - E)^2 \right\rangle_\beta = - \frac{\partial E}{\partial \beta} = N k_B T^2 c_v, \quad (15)$$

is a particular case of Eq. 14 with $f_j = \mathcal{H} = \omega$ and $\lambda_j = \beta$.

Notice that the FDT is also valid in the case of discrete states but continuous parameters. For instance, in the case of the Poisson distribution,

$$P(k|\lambda) = \frac{\lambda^k \exp(-\lambda)}{k!}, \quad (16)$$

it is possible to write the FDT as

$$\frac{\partial}{\partial \lambda} \langle \omega(n, \lambda) \rangle_\lambda = \left\langle \frac{\partial \omega}{\partial \lambda} \right\rangle_\lambda + \left\langle \omega \left(1 - \frac{k}{\lambda} \right) \right\rangle_\lambda, \quad (17)$$

from which it quickly follows (using $\omega = \lambda$) that $\lambda = \langle k \rangle_\lambda$.

Direct applications of the CVT in Thermodynamics

We may consider the equipartition theorem as the original application of the CVT, in much the same way as the canonical ensemble was the original application of MaxEnt. For a classical system with Hamiltonian $\mathcal{H}(\Gamma)$ in the canonical ensemble, the CVT takes the form

$$\left\langle \frac{\partial \omega}{\partial \Gamma_i} \right\rangle_\beta = \beta \left\langle \omega \frac{\partial \mathcal{H}}{\partial \Gamma_i} \right\rangle_\beta \quad (18)$$

and the equipartition theorem [1] follows by choosing $\omega(\Gamma) = \Gamma_j$,

$$\beta \left\langle \Gamma_j \frac{\partial \mathcal{H}}{\partial \Gamma_i} \right\rangle_\beta = \delta_{ij}. \quad (19)$$

From this identity it follows that, if we choose Γ_j as any quadratic degree of freedom, i.e.,

$$\mathcal{H} = \alpha_j \Gamma_j^2 + \mathcal{H}'(\Gamma_1, \dots, \Gamma_{j-1}, \Gamma_{j+1}, \dots, \Gamma_N) \quad (20)$$

then $\Gamma_j \partial_j \mathcal{H} = 2\alpha_j \Gamma_j^2$ and therefore $\left\langle \alpha_j \Gamma_j^2 \right\rangle_\beta = k_B T/2$, which is the standard statement of the equipartition theorem,

“Each quadratic degree of freedom in the Hamiltonian contributes exactly $k_B T/2$ to the average energy”.

A generalization is the following expression for the inverse temperature, first reported by Rugh [8] and Rickayzen [9] in the context of the microcanonical ensemble,

$$\beta = \left\langle \nabla \cdot \left[\frac{\mathbf{v}(\Gamma)}{\mathbf{v}(\Gamma) \cdot \nabla \mathcal{H}} \right] \right\rangle_\beta. \quad (21)$$

which can be obtained from the vector form of the Equation 18,

$$\left\langle \nabla \cdot \omega \right\rangle_\beta = \beta \left\langle \omega \cdot \nabla \mathcal{H} \right\rangle_\beta \quad (22)$$

by choosing $\omega = \mathbf{v}/(\mathbf{v} \cdot \nabla \mathcal{H})$. The reason this result works for the canonical ensemble even when the original derivation was in the microcanonical ensemble is, besides the ensemble equivalence in the thermodynamic limit, the fact that this construction can be extended to any ensemble in Statistical Mechanics where the probability distribution of the microstates $P(\Gamma|S) = \rho(\mathcal{H}(\Gamma))$ is dependent only on \mathcal{H} .

In order to see this, we use the chain rule,

$$\frac{\partial}{\partial \Gamma_i} \ln \rho(\mathcal{H}) = \left[\frac{\partial}{\partial E} \ln \rho(E) \right]_{E=\mathcal{H}} \frac{\partial \mathcal{H}}{\partial \Gamma_i} \quad (23)$$

and substitute in Equation 10, leading to

$$\left\langle \frac{\partial \omega}{\partial \Gamma_i} \right\rangle_S = \left\langle \hat{\beta}(\mathcal{H}) \omega \frac{\partial \mathcal{H}}{\partial \Gamma_i} \right\rangle_S, \quad (24)$$

where we have defined the quantity

$$\hat{\beta}(E) = -\frac{\partial}{\partial E} \ln \rho(E) \quad (25)$$

as an analog of the Lagrange multiplier β . Now the same substitution $\omega = \mathbf{v}/(\mathbf{v} \cdot \nabla \mathcal{H})$ as before produces

$$\left\langle \hat{\beta}(\mathcal{H}) \right\rangle_S = \left\langle \nabla \cdot \left[\frac{\mathbf{v}}{\mathbf{v} \cdot \nabla \mathcal{H}} \right] \right\rangle_S. \quad (26)$$

which shows that the expectation in the right-hand side is independent of the choice of \mathbf{v} . The canonical ensemble is the trivial case where $\hat{\beta}$ reduces to the constant β .

Thus, the CVT reveals a family of “temperature estimators” which can be used to measure the inverse temperature β in different ensembles.

Dynamical systems and the functional version of CVT

Soon after the proposal by Jaynes of the Maximum Entropy principle, it was realized [10, 11] that the same principle should apply to inference over dynamical systems. The application to time-dependent systems became known as the Maximum Caliber³ principle.

A brief and simplified formulation of this version of the principle is as follows. Consider an unknown trajectory $x(t)$ of a one-dimensional system where we instantly know the joint probability density $P(x(t) = X, \dot{x}(t) = V) = P(X, V|t)$ of position and velocity. The most unbiased probability **functional** for the possible trajectories $x(t)$ is, according to the Maximum Caliber principle [12, 13], the one that maximizes the path entropy

$$S[P] = - \int Dx() P[x()] \ln P[x()] \quad (27)$$

subjected to the constraints

$$\left\langle \delta(x(t) - X) \delta(\dot{x}(t) - V) \right\rangle = P(X, V|t). \quad (28)$$

The solution is of the form,

$$P[x()|\mu()] = \frac{1}{\eta[\mu()]} \exp \left(- \int dX dV \int_0^\tau dt \mu(X, V, t) \delta(x(t) - X) \delta(\dot{x}(t) - V) \right), \quad (29)$$

which after explicitly integrating the delta functions, gives

$$P[x()|\alpha] = \frac{1}{\eta(\alpha)} \exp \left(- \int_0^\tau dt \mu(x(t), \dot{x}(t), t) \right) = \frac{1}{\eta(\alpha)} \exp \left(- \frac{1}{\alpha} A[x()] \right), \quad (30)$$

where

$$A[x()] = \int_0^\tau dt \mathcal{L}(x(t), \dot{x}(t), t) \quad (31)$$

has the form of the action in classical mechanics with Lagrangian $\mathcal{L}(x, \dot{x}; t) = \alpha \mu(x, \dot{x}, t)$, and we have extracted a global factor $1/\alpha$ with units of action from the original Lagrange multiplier μ .

By discretizing time, i.e., considering the trajectory $x(t)$ as a sequence of discrete time snapshots $x = (x_1, \dots, x_N)$ separated by a time step δt so that $\tau = N\Delta t$, we see that the Maximum Caliber problem can be mapped into a MaxEnt problem with N degrees of freedom. It follows that, by the identification [14]

$$\frac{1}{\Delta t} \frac{\partial}{\partial x_i} \equiv \frac{\delta}{\delta x(t)} \quad (32)$$

the CVT takes the *functional* form

$$\left\langle \frac{\delta W[x()]}{\delta x(t')} \right\rangle_I = \frac{1}{\alpha} \left\langle W[x()] \frac{\delta A[x()]}{\delta x(t')} \right\rangle_I. \quad (33)$$

where now W is any arbitrary, differentiable *functional*. Taking $W = 1$ we see that the Euler-Lagrange equation for the Lagrangian of the problem is fulfilled in expectation,

$$\left\langle \frac{\delta A[x()]}{\delta x(t)} \right\rangle_I = \left\langle \frac{\partial \mathcal{L}}{\partial x} - \frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{x}} \right) \right\rangle_I = 0. \quad (34)$$

which is, of course, Equation 11. This result for the particular case of a one-dimensional particle in a potential was obtained, in discretized form, in Ref. [12], and shows that the ensemble of trajectories follows Newton's law of motion,

$$\left\langle \frac{dp}{dt} \right\rangle = \left\langle - \frac{d\Phi(x)}{dt} \right\rangle, \quad (35)$$

in expectation, where $p = m\dot{x}$ is the momentum and $\Phi(x)$ the potential energy.

³Because the entropy of paths is analogous to the cross section of a tube, or the caliber of a gun barrel.

CVT in Bayesian parameter estimation

Consider a statistical model $P(x|\theta)$ for which we know n observations x_1, \dots, x_n , denoted collectively by D . The posterior distribution for the parameter θ is given by Bayes' theorem,

$$P(\theta|D) = \frac{P(\theta|\mathcal{I}_0)P(D|\theta)}{P(D|\mathcal{I}_0)}, \quad (36)$$

For the case where the n observations are statistically independent, we have

$$P(D|\theta) = \prod_{i=1}^n P(x_i|\theta) \quad (37)$$

and therefore

$$P(\theta|D) = \frac{1}{\eta} P(\theta|\mathcal{I}_0) \exp\left(\sum_{i=1}^n \ln P(x_i|\theta)\right), \quad (38)$$

where $\eta = P(D|\mathcal{I}_0)$. Now the x_i are given (could be taken as the fixed parameters of the posterior model) and the single variable θ is the sole degree of freedom of our problem. The CVT for this posterior distribution is

$$\left\langle \frac{d\omega(\theta)}{d\theta} \right\rangle_D = - \left\langle \omega(\theta) \frac{\partial}{\partial \theta} \ln P(\theta|D) \right\rangle_D = - \left\langle \omega(\theta) \left\{ \frac{\partial}{\partial \theta} \ln P(\theta|\mathcal{I}_0) + n \frac{\partial}{\partial \theta} \ln P(x|\theta) \right\} \right\rangle_D. \quad (39)$$

where $\bar{a} = \frac{1}{n} \sum_{i=1}^n a(x_i)$ is the arithmetic mean over the data D . If our original model $P(x|\theta)$ is a MaxEnt model,

$$P(x|\theta) = \frac{1}{Z(\theta)} \exp(-\theta f(x)), \quad (40)$$

then Equation 39 simplifies to

$$\left\langle \frac{d\omega(\theta)}{d\theta} \right\rangle_D + \left\langle \omega(\theta) \frac{\partial}{\partial \theta} \ln P(\theta|\mathcal{I}_0) \right\rangle_D = n \left\langle \omega(\theta) [\bar{f} - F(\theta)] \right\rangle_D, \quad (41)$$

where we have used

$$-\frac{d}{d\theta} \ln Z(\theta) = \langle f \rangle_\theta = F(\theta). \quad (42)$$

For a large number of observations n , the effect of the prior vanishes and we simply get

$$\left\langle \frac{d\omega(\theta)}{d\theta} \right\rangle_D = n \left\langle \omega(\theta) [\bar{f} - F(\theta)] \right\rangle_D. \quad (43)$$

From this last identity it follows that

$$\lim_{n \rightarrow \infty} \bar{f} \rightarrow \langle F(\theta) \rangle_D = \langle f(x) \rangle_D \quad (44)$$

$$\langle (\Delta F(\theta))^2 \rangle_D = -\frac{1}{n} \left\langle \frac{dF(\theta)}{d\theta} \right\rangle_D \quad (45)$$

by using $\omega = 1$ and $\omega = \bar{f} - F(\theta)$ respectively. Together they imply the law of large numbers: if $dF(\theta)/d\theta$ is finite, then $\bar{f} = \langle f(x) \rangle_D$ when $n \rightarrow \infty$. Note that

$$\frac{dF(\theta)}{d\theta} = - \left\langle (\Delta f(x))^2 \right\rangle_\theta \quad (46)$$

by the theorem in Equation 14.

CONCLUSIONS

We have shown the versatility of the divergence theorem applied to probability densities of continuous systems, in a form we have called the conjugate variables theorem (CVT). The same relation that provides the meaning of temperature associated with average kinetic energy in Statistical Mechanics can be used to obtain properties of non-equilibrium systems in terms of path averages and as a shortcut in Bayesian parameter estimation. The conjugate variables theorem (Eq. 10) together with the fluctuation theorem in Eq. 13, provide a new set of tools to approach inference problems for continuous variables/parameters.

ACKNOWLEDGMENTS

SD acknowledges support from FONDECYT grant 1140514.

REFERENCES

- [1] K. Huang, *Statistical Mechanics* (Wiley, 1987).
- [2] E. T. Jaynes, *Physical Review* **106**, 620–630 (1957).
- [3] A. Caticha and A. Giffin, *AIP Conf. Proc.* **872**, p. 31 (2006).
- [4] S. Davis and G. Gutiérrez, *Phys. Rev. E* **86**, p. 051136 (2012).
- [5] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley and Sons, 2006).
- [6] E. T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, 2003).
- [7] H. Callen, *Thermodynamics and an Introduction to Thermostatistics* (Wiley, 1985).
- [8] H. H. Rugh, *Phys. Rev. Lett.* **78**, 772–774 (1997).
- [9] G. Rickayzen and J. G. Powles, *J. Chem. Phys.* **114**, p. 4333 (2001).
- [10] A. A. Filyukov and V. Y. Karpov, *Inzh.-Fiz. Zh.* **13**, p. 624 (1967).
- [11] E. T. Jaynes, *Ann. Rev. Phys. Chem.* **31**, 579–601 (1980).
- [12] D. González, S. Davis, and G. Gutiérrez, *Found. Phys.* **44**, p. 923 (2014).
- [13] S. Davis and D. González, *J. Phys. A: Math. Theor.* **48**, p. 425003 (2015).
- [14] I. M. Gelfand and S. V. Fomin, *Calculus of Variations* (Dover Publications, 2000).